

⑫ 公開特許公報(A) 平2-205970

⑤Int.Cl.⁵

識別記号

庁内整理番号

⑬公開 平成2年(1990)8月15日

G 06 F 15/40

5 0 0 T

7313-5B

審査請求 未請求 請求項の数 14 (全12頁)

⑭発明の名称 データ記憶及び検索方法及びスキヤナ

⑮特 願 平1-11752

⑯出 願 平1(1989)1月20日

⑰発 明 者	フオーブス・ジェイ・ バーコブスキ	カナダ国、エヌ2ジェイ・2シー8、オンタリオ、ウオー タールー、マーガレット・アベニュー・サウス 65
⑱発 明 者	マーク・シンクレア・ クレブス	カナダ国、エム6ジー・1ブイ5、オンタリオ、トロ ント、ガーネット・アベニュー 46
⑲出 願 人	フオーブス・ジェイ・ バーコブスキ	カナダ国、エヌ2ジェイ・2シー8、オンタリオ、ウオー タールー、マーガレット・アベニュー・サウス 65
⑲出 願 人	マーク・シンクレア・ クレブス	カナダ国、エム6ジー・1ブイ5、オンタリオ、トロ ント、ガーネット・アベニュー46
⑳代 理 人	弁理士 鈴江 武彦	外3名

明 細 書

1. 発明の名称

データ記憶及び検索方法及びスキヤナ

2. 特許請求の範囲

(1) (a) データ記憶装置にデータベースを記憶する工程と、

(b) サブセットに分割されるシグナチャーファイルを上記データベース用に作成する工程、上記ファイルの作成時に特定のサブセットにワードシグナチャーをマッピングする工程、及び上記データ記憶手段の上記シグナチャーファイルを記憶する工程と、

(c) ワードシグナチャーを走査する工程、及び特定のサブセットに上記ワードシグナチャーを記憶するために使用された同じマッピング情報を使用することによるクイリーキーワードに応じた上記データベースから対応データを検索する工程と

を具備し、上記データベースの情報を記憶すると共に検索するもので、データ処理手段と、メモリ

手段と、データレコードを格納することが可能なデータ記憶手段を有するコンピュータシステムを動作することを特徴とするデータ記憶及び検索方法。

(2) 一つのクイリーキーワードに対する応答としてワードシグナチャーとの一致をとる場合、上記クイリーキーワード用の物理ワードシグナチャーを含むサブセットのみを走査するシステムを有する工程を更に具備することを特徴とする特許請求の範囲第1項記載のデータ記憶及び検索方法。

(3) 特定の文書に対応するシグナチャーファイル作成に当たって上記文書の全ての共通ワードを無視する工程と、上記文書で残ったワードの各々に対応した論理ワードシグナチャーを計算する工程と、論理ワードシグナチャーがハッシュ値として計算された場合は上記文書中の何れか重複した論理ワードシグナチャーを除去する工程を更に具備することを特徴とする特許請求の範囲第2項記載のデータ記憶及び検索方法。

(4) 上記シグナチャーファイルの作成に当たっ

て上記シグナチャーファイルを多数のサブセットに分割する工程と、各論理ワードシグナチャーが二つの成分から構成されるように生成する工程と、同一のサブセット指定フィールドを有する全ての論理ワードシグナチャーを同じサブセットにマッピングする工程と、上記論理ワードシグナチャーのサブセット指定フィールドに連結している物理ワードシグナチャーの成分のサブセット部分の記憶において結果的に生じる工程とを更に含むことを特徴とする特許請求の範囲第3項記載のデータ記憶及び検索方法。

(5) 上記サブセットの作成に当たって特定の物理ワードシグナチャー若しくは物理ワードシグナチャーのグループが誘導された元となる上記文書に文書識別記号を与え、上記同一の文書から誘導された上記物理ワードシグナチャー若しくは物理ワードシグナチャーのグループに加える工程を更に具備することを特徴とする特許請求の範囲第4項記載のデータ記憶及び検索方法。

(6) 複数のビットから成る整数値のワードをマ

チャーファイルの大きさを増大する工程を更に具備することを特徴とする特許請求の範囲第1項記載のデータ記憶及び検索方法。

(9) 上記論理ワードシグナチャーはその対応キーワードに関連されるべくものであり、上記論理ワードシグナチャーと上記対応キーワードとの関係はキーワード辞書で保持されるものであり、全てのサブセットの長さを等しく維持するように上記サブセット指定フィールドを選択すると共に上記論理ワードシグナチャーと上記対応キーワード間のマッピングが一对一になるように上記物理ワードシグナチャーを選択する工程とを更に含むことを特徴とする特許請求の範囲第8項記載のデータ記憶及び検索方法。

(10) 上記データ記憶装置はディスクであり、予め明記された物理ワードシグナチャーを求めて上記ディスクから読出されるデータストリームを走査するスキナに関連した上記データ処理手段を使用する一つ以上のサブセットを取出す工程を更に具備することを特徴とする特許請求の範囲第3

項記載のデータ記憶及び検索方法。

(7) 上記論理ワードシグナチャーは m ビット長から成るもので上記 m は8乃至32までの範囲であり、且つ多数のサブセットから一つを選択するサブセット指定フィールドに連絡される n ビットの物理ワードシグナチャーを有して上記 n は8乃至20の範囲であることを特徴とする特許請求の範囲第6項記載のデータ記憶及び検索方法。

(8) 上記サブセットは大きさが増大された上記データベースとして上記サブセットの大きさの増大を許容するために連続するサブセット間に任意のスペースを有するように上記データ記憶装置に配置され、その終端に付加的な情報を追加することによって上記データベースの大きさと、一つ以上のサブセットの終端で付加的なワードシグナチャーグループを追加することによって上記シグナ

項記載のデータ記憶及び検索方法。

(11) 上記各 n ビットの物理ワードシグナチャーはRAMに於ける2の n 乗に相当するロケーションのアドレスとして使用されるもので上記 n は8乃至20の範囲の整数であり、各ロケーションは1ビットとを保有し、上記スキナは上記アドレスされたRAMのビットが上記クイリーに対し無関係なワードシグナチャーを指示したときは物理ワードシグナチャーを無視し、いかなる物理ワードシグナチャーでもFIFOに将来参考用として納めるように制御し、上記アドレスされたRAMビットロケーションが相補的な値を有するときは上記スキナは求めんとする物理ワードシグナチャーを取出すことを特徴とする特許請求の範囲第10項記載のデータ記憶及び検索方法。

(12) 上記シグナチャーファイルの作成時に各サブセットを作成するのに一連のグループを生成する工程を含み、各グループは一連の物理ワードシグナチャーを有し、各グループは文書識別記号の表示値により終結され、各文書識別記号の表示値

は高、中及び低位フィールドを有し、上記サブセットの第1のグループの上記文書識別記号の表示値は実際に高、中及び低位フィールドを有してそれ以後のグループからの文書識別記号の表示値は常に低位フィールドを有するが、中位フィールド若しくは高位フィールドに関しては直前のグループに表示された上記文書識別記号からこれらのフィールドの変化を反映するのに必要なとき中位フィールド若しくは高位フィールドのみのフィールドが低位フィールドに加えて使用され、上記文書識別記号は数の増加する方向に配列されることを特徴とする特許請求の範囲第11項記載のデータ記憶及び検索方法。

(4) 上記データ記憶手段からの情報を検索する場合、上記データ記憶手段からデータストリームを走査するためにスキャナを使用する工程と、文書識別記号の全ての表示値を押下するスキャナは上記FIFOに送込む高位フィールドを有し、上記スキャナは将来の参考のためにレジスタに上記データストリームの最後に遭遇した中位のフィールド

具備し、上記入力部は上記データ記憶手段から情報を受取り、上記制御部は上記情報を調べて上記メモリの上記アドレスラインに全てのワードシグナチャーを送り、上記メモリは上記入力部で特定のワードシグナチャーが上記クイリーキーワード用のワードシグナチャーと一致するか否かを決定するために上記制御部に情報を提供し、一致が生じた場合は、

(a) 上記制御部は上記ワードシグナチャーを上記FIFOバッファに送り、一致が生じたことを記憶しており、この一致したワードシグナチャーの次に位置する文書識別記号を上記FIFOバッファに送り、その後上記データ記憶手段から受取った次のワードシグナチャーを処理するように進行し、

一致が生じない場合は、

(b) 上記制御部は上記ワードシグナチャーを無視してその後上記データ記憶手段から受取った次のワードシグナチャーを処理するように進行し、

を与え、上記最後に遭遇した中位のフィールドは上記クイリーキーワードから誘導された上記物理ワードシグナチャーとその特定のグループ内の物理ワードシグナチャー間の一致を上記RAMが指示した場合のみ特定のグループを終結される上記低位フィールドとに送込まれることを特徴とする特許請求の範囲第12項記載のデータ記憶及び検索方法。

(4) データベースに於ける情報を記憶すると共に検索するためにコンピュータシステムの使用のためのものであって、上記コンピュータシステムはデータ記憶装置の上記データを記憶するための手段と共に、データ処理手段と、メモリ手段と、データ記憶手段と、上記データベース用の文書識別記号と共にワードシグナチャーファイルを作成する手段と、ワードシグナチャー、文書識別記号及びスキャナを使用してクイリーキーワードに応じたデータベースから対応データを検索する手段とを有し、入力部と、出力部と、制御部と、アドレスライン及びFIFOバッファから成るメモリとを

上記制御部は実質的には数個のクイリーキーワードを並行に処理することが可能なことを特徴とするスキャナ。

3.発明の詳細な説明

〔産業上の利用分野〕

本発明はデータ検索システムに関し、特に細分化されたシグナチャーファイルを用いてデータベースにデータを蓄積し又はデータベース上のデータを検索する方法及びこのシステムを使用するスキャナに関する。

〔従来の技術〕

一般に数種のデータ蓄積検索システムが知られている。近來、データベースが大型化し、その使用頻度が上昇するに及んで、データを正確にしかも最少の時間で蓄積したり検索したりする方法が益々重要視されて来ている。更に又、大規模な変更を必要としないデータ追加の方法も重要である。

従来のデータベースに於いては情報は周到に用意された索引形態として高度に組織化され、たとえばディスク等の蓄積媒体に格納されている。今、

仮にある特定の一部のデータを検索しようとする
と、従来のシステムはこのデータの所在を突き止
めるのにこの索引情報を用いるが、この要求され
るデータは大量の同類のデータの山に埋もれてい
る場合が多い。索引情報は余分のファイルを必要
とし、これに蓄えられる。索引を作るには様々な
方法がある。シグナチャーファイルを用いる事は
その一つの方法であり、逆見出しファイルを用い
るのがもう一つの方法である。後者の方法は度々
用いられるものであり、検索時間が少なくて済む
が、次に述べる如く二つの重大な欠点を持っている。

(i) 逆見出しファイルは非常に大きく、テキス
トファイルの20%から100%の大きさに相当
する。

(ii) データベースへの新情報の追加に際しては
逆見出しファイルを変更する必要がある。この変
更に要する時間が大変長くなる。何故ならば高度
に組織化されたファイルの性格上、ファイルの大
部分を変更する必要があるからである。

には一度しか情報を書き込むことができないので
データ更新に問題がある。したがって、新情報は、
通常現在の情報の隣に書き込むのが望ましいもの
とされるが、その場所がスペースとして残されて
いないので不可能である。故に、逆見出しファイ
ルのような索引形態では、この様な蓄積媒体のデ
ータの非消去性の為に変更する事は、通常出来な
い。新しいファイルを別途にディスクの新しい領
域に作成すれば変更可能であるが、これはディス
クスペースの無駄が大きく非能率的である。

シグナチャーファイルを用いる場合、データベ
ースに追加される情報に対応してシグナチャーフ
ァイルに追記が施される。したがって、ファイル
そのものの量は増大するが現存するシグナチャー
ファイルには変更が加えられないから、光ディス
クを用いたシステムに於いても大変有効である。

多くの場合、検索システムは使用者が予めキー
ワードとしてクイリー中に明記した一つ以上のワ
ードを含む一つ以上の文書を検索することによって、
この使用者クイリーに対して応答する。これを連

シグナチャーファイルの場合は迅速な変更が可
能であり、アクセスタイムが大変遅くなるという
理由から索引形式は適用されない。アクセスタイ
ムが遅くなるのは、全てのシグナチャーファイル
を走査する為であり、ディスクからの転送所要時
間が大変長くかかるからである。本発明ではシグ
ナチャーファイルを用いるものであるが、逆見出
しファイルのスピードに対抗し得る工夫がなされ
ている。つまり、全てのファイルを走査する代わ
りに、サブセットのみを走査する。これにより、
アクセスタイムを著しく短少する事が可能となり、
システムを注意深く設計すれば、変更時間もかな
り低く保つ事が出来る。

〔発明が解決しようとする課題〕

勿論、全てのデータベースを走査して(索引フ
ァイルを使用せずに)データを検索する事は可能
であるが、それには大変長時間を要し、従って問
題にならない程高価になる。

データ蓄積装置として光ディスクが用いられる
場合、現在の開発段階では光ディスクの特定領域

成する為、データベース中の要求される情報の所
在を指示する検索機能が用いられる。これらのキ
ーワードを含む文書の所在位置を求める為に文書
識別記号リストを作成する事により、クイリー応
答用ソフトウェアに関連して作用するこの索引機
能はクイリーの要求を満たすべく最終的な文書リ
ストを決定する。

データベース中のデータの所在を突き止める為
にシグナチャーファイルを用いる事は、既存の技
術である。シグナチャーファイルとは、即ちデー
タベース中の情報の凝縮されたものである。これ
は、データベース中の文書の各々の明確なワード
をワードシグナチャーによって表わす事によつて
達成される。ある特定のクイリーキーワードがシ
ステムに提示されると、システムはそれに対応し
たデータベース中のワードに関連したワードシグ
ナチャーを引き出す。この種のシステムは、この
ようにして連続走査の方法を用いる事により、全
てのシグナチャーファイルを検索し、そのキー
ワードを含んだデータベース中の全ての文書を検し

出す事が出来る。これは即ち、シグナチャーファイルの何れかのワードシグナチャーが、それら自身が誘導されたキーワードを含む文書の文書識別記号によって可能となる。従って、走査プロセスの進行期間中に、シグナチャーファイル中のワードシグナチャーがクイリーキーワードから誘導されたワードシグナチャーと一致した時、システムはそのクイリーに関連した文書の識別を保持しておく為、ワードシグナチャーに付属した文書識別記号を取得する。これらのシステムは、もし各々のクイリーキーワードごとに全てのシグナチャーファイルを検索するのであれば、やはり長時間を要する事になる。

光ディスクは、データベースを格納するには最も経済的な手段である。しかしながら光ディスクのシークタイムは、磁気ハードディスクに比べて通常4倍から30倍も長い。逆見出しファイル方式を用いてデータベースが検索された場合、システムはおそらく数回に渡って索引構造を探索する事になり、毎回の探索各にディスクシーク、即ち

ディスクアームの動きを必要とする。より高価な磁気ハードディスクを用いれば、時間的要求は満たされるであろうが、光ディスクを使用する場合には極端に望ましくないものとなる。

本発明の目的は以上の事柄を考慮してなされたもので、比較的高速なデータの記憶及び検索を行う^{方法}を提供する事である。

〔課題を解決するための手段〕

本発明は、シグナチャーファイルの検索が一回のみで十分であり、且つ特定のクイリーキーワードに対する応答としての検索が単にシグナチャーファイルの一部のみで済ませ得ることを特徴としている。

〔作用〕

本発明によれば、データ処理手段と、メモリ手段と、データレコードを含むデータ記憶手段を有するコンピュータシステムに於いて、これを運用してデータベース上に情報を記憶し又データベース上で情報を検索するコンピュータシステムの動作方法は次の通りである。

(1) データ記憶装置にデータベースを記憶する工程。

(2) 複数のサブセットに分割されるシグナチャーファイルをデータベース用に作成する工程と、ファイル作成中に特定のサブセットに対応したワードシグナチャーをマッピングする工程と、上記シグナチャーファイルサブセットを上記データ記憶装置に記憶する工程。

(3) ワードシグナチャーを走査する工程及び特定のサブセットへワードシグナチャーを蓄積するために使用された同じマッピング情報を用いる事により上記データベースからクイリーキーワードに応じて対応するデータを検索する工程。

スキナは、データベース上に情報を記憶すると共に検索する為にコンピュータシステムの使用を提供する。コンピュータシステムは、データ処理手段と、メモリ手段と、データ記憶手段を有する。これらはデータベース用として文書識別記号とともにワードシグナチャーファイルを作成する手段と、スキナを用いてクイリーキーワードに

応じてデータベースからワードシグナチャー、文書識別記号及びそれに対応したデータを検索する手段とを有している。スキナは入力部、出力部、制御部、アドレスラインを有したメモリ及びFIFOバッファを有する。入力部は上記データ記憶手段からの情報を受けるために接続されている。制御部は上記情報を調べて上記メモリのアドレスラインに全てのワードシグナチャーを送る。メモリは制御部へのクイリーキーワード用のワードシグナチャーと入力部に現れたある特定のワードシグナチャーとが一致するか否かを決定し得る情報を提供する。もし一致すると、制御部はそのワードシグナチャーをFIFOバッファに送り、且つ一致の発生を記憶する。制御部は更に一致したワードシグナチャーの次に位置する文書識別記号をFIFOバッファに送る。制御部はこの後、順次データ記憶手段等から受け取る次のワードシグナチャーを処理してゆく。もし一致が認められなければ、制御部はそのワードシグナチャーを無視してデータ記憶手段から受け取る次のワードシグナチャーを処理

する。従って、制御部は実質的に複数のクイリーキーワードを平行して処理する事ができる。

〔実施例〕

以下図面を参照して、本発明の実施例を説明する。

第1図により、本発明に従ったデータ記憶及び検索システムを用いれば、多数の使用が同時に夫々異なったクイリーキーワードを用いて夫々異なったデータの検索を可能にするため設計されるという事が判る。更に本システムは、複数の光ディスクユニットを有して夫々のユニットにデータベースが蓄えられると共にデータを検索することができる。

本発明のこの実施例に於いて、スキャナモジュールは同時に4096個のワードシグナチャーの検索が可能である。データベースのシグナチャーファイルに於いて一度クイリーキーワードのワードシグナチャーが検出されると、それに対応したキーワードを含む全ての文書の識別記号が候補文書リストとして収集される。もし望ましいもので

成されデータ記憶装置又は光ディスクに記憶される。シグナチャーファイルは、一連の整数を有し(固定長のビット列)、各々の整数はデータベースの主テキストに含まれた重要なワードの実際のワードシグナチャーを表わしている。特定の文書のシグナチャーファイルが作成される時には、次の三つのステップをもってなされる。

1) 共通ワードは停止ワードのリストを用いて除去される。

2) 上記文書の明確な残りのワードごとに論理ワードシグナチャーが計算される。これは単にmビット長の整数値にワード(文字列)をマップするハッシュ機能であってもよい。ここでmは8から32までの整数とする。好ましくは、各論理ワードシグナチャーは二つの要素を有するために発生され、nビットの物理ワードシグナチャーがサブセット指定フィールドに連鎖的に連がれる多数のサブセットから一つを選択する。この場合、nは8から20までの数である。

3) 重複ワードシグナチャーはハッシュ機能を用

あれば、特定の文書がクイリーの要求を満たすか否かを判定するこのリストを処理するためにソフトウェアを作成する事も可能である。使用者に必要な文書の所在を知らせたら、必要に応じて実際の文書を調べる為に検査すれば良い。

データ処理手段は、プロセッサボードに関連して作動するハードウェアとしてのスキャナモジュールと、小型のシャーシに納まった各種の入出力モジュールを有する。このユニットは、キーワード検出に適切であり使用者のワークステーションと全体のテキスト及びシグナチャーファイルの全てを保持するために使用されるデータ記憶手段(光ディスク又は磁気ディスク)の両者と連絡している。ワークステーションコンピュータは、クイリー受入れ及びクイリー分析、そして走査用コンピュータとの連絡に関する全ての処理を行う。走査用コンピュータは在来型のシリアルライン例えばRS-232リンク又はETHERNET(商標)等の高帯域幅機能を通してワークステーションと連絡する。

シグナチャーファイルは、データベースから作

いて計算する事により防止される。

もし文書ワードとそれに対応した論理ワードシグナチャーとが一对一のマッピングでしみ込めさせるような方法で割当てられると、このステップは省略される。

停止ワードとは、一般にクイリーの要求が満たされたときに文書の違いを見分けるのに寄与しないワードの事である。これらは通常接続詞とが冠詞等である。例えば共通ワードとしては、"a" "the" "when" "where" "henceforth" 等である。停止ワードのリストはシステムの記憶領域に覚えさせてあり、システムは自動的にこれらの停止ワードを見逃ごしてワードシグナチャーの作成を行わない。

作成された各々のシグナチャーファイルサブセットは、対応するテキスト文書と同じ順序に現れる一連の文書シグナチャーグループである。各々の文書シグナチャーグループは、それに対応したテキスト文書のワードから誘導された一連のワードシグナチャーから成っている。一つの文書に拘

わるシグナチャーグループの最後に記入されるものは、その文書を代表する認識記号の表示である。使用者のクイリーに解答が与えられる為には、先ず使用者のクイリーから発せられたクイリーキーワードが論理ワードシグナチャーに変換され、次の物理レベルのワードシグナチャーとの一致を見る為にシグナチャーファイルが走査される。一致がとれると、それに対応したテキスト文書の所在を定める事が出来るのでシグナチャーグループの最後に記述された文書記号が抽出される。

ハッシュ機能を使って論理ワードシグナチャーを作成する事は可能である。この方法は、非常に少量の記憶容量で早く作成する事が出来る。しかしながら、以下に述べる如き欠点を伴う。すなわち、特定のワードシグナチャーのハッシュエンコーディングでは必ずしも一対一のマッピングが保証されず、従って異なったワードが同じ論理ワードシグナチャーとマップする事があり得る。従って走査実行中にこの複数回のマップは、使用者には全く必要としない無関係の文書を検索してしまう結果を

添付される場合は、シグナチャーファイルも同様に文書シグナチャーグループの形式で追加ワードシグナチャーを添付する必要がある。このようにシグナチャーファイルが簡単に最新の状態に更新できる事は、シグナチャーファイル使用の確固たる利点とされる。

シグナチャーファイルの走査は全体のデータベースの走査に較べて遙かに速いが、シグナチャーファイルの走査も場合によってはかなりの時間を要し、データ記憶装置に光ディスクが用いられる時は特に長時間を要する。たとえば、データベースのサイズが700メガバイトある時、シグナチャーファイルは約35メガバイトの大きさになる。光ディスクからデータ処理手段へのデータストリームは、通常毎秒1メガバイト位であるからシグナチャーファイルの何れの走査には少なくとも35秒を要する。

更新の簡易さの利点を維持しながら光ディスクの走査時間を最短にする為、シグナチャーファイルは無理のない程度の数のサブセット(例えば

招く事がある。これは事前に検知する事が可能であり、使用者に渡る前にソフトウェアを組んで検査し、修正しておく事が可能である。ワードシグナチャーが十分長ければ、この様な間違いは減少する事が出来る。シグナチャーファイルは、ハッシュ機能を用いる事により数ビットから成る整数値に各々のワードをマップし、文書にポインターを誘導する為の文書識別記号を作成する事により生成される。

ワードシグナチャーは、通常それが表わしている個々のテキストワードより遙かに短い。更に共通ワード及び重複シグナチャーは取り除かれているので、シグナチャーファイルはそれ自身が誘導された元のデータベースよりかなり短くなっている。通常、シグナチャーファイルの長さはデータベースの5%から30%の長さである。もしシグナチャーファイルが文書のスタートの概要前に置かれる抜粋用として関連語や同義語等を作成する用意を持っている場合、シグナチャーファイルは多少大きくなる。もしデータベースに追加文書が

256個のサブセット)に分割される。シグナチャーファイル作成時、ワードシグナチャーは特定のサブセットにマップされ、クイリーキーワードに応答してワードシグナチャーをサブセットが走査される時その同じマッピング情報が用いられる。

種々のワードシグナチャーの形式が可能である。以下に述べる形式は一つの例と見なされる。15ビットの物理ワードシグナチャーがステアリングビットとして最上位ビットでディスク上に2つのカンセクティブバイトとして記憶され、このステアリングビットの一つの設定値は、相補的な設定値がこの16ビットワードが文書の識別記号を与えることをする一方、この16ビットのワードが物理ワードシグナチャーを含むことを意味している。もしディスクが256のサブセットに分割されている場合は、ハッシュングによって生成された論理ワードシグナチャーの実質長は23ビットとなり、従って物理ワードシグナチャーが単に15ビットしかないにも拘わらず間違いを大変低く保つ事が出来る。

タイリーキーワードに応じて走査時間は全体のシグナチャーファイルのほんの一部を走査するためにだけ必要なもので、かなりの減少となる。細分化されていないシグナチャーファイルに較べシグナチャーファイルの細分化は、ファイルによるスペースの総使用量の増加につながる。何故ならば、ある文書の全てのワードシグナチャーグループは、多数のサブセットに分散された為にこの全てのワードシグナチャーの一部を有する全てのサブセットに文書識別記号を必要とするからである。

いま、各々2バイトからなるワードシグナチャーが2バイトの文書ポイントにより従われると、この最悪の場合、ワードシグナチャーは4バイトを占める事になる。シグナチャーファイルが細分化されていない場合、このファイルへの一単位の情報に文書番号を除外しても通常24ビット、3バイトのシグナチャーとなる。従って、多数のサブセットに分割されたシグナチャーファイルの総量は、細分化されていないシグナチャーファイルの総量の三分の一以下の増加に止まる。

る。スキナは予め決められた物理ワードシグナチャーを搜索する。

どのサブセットに特定のワードシグナチャーを持たせるかについては、システムがサブセットの増大が均一になる様に調整している。実際に、データベースの初期の大きさがかなり大きい場合、サブセットがほぼ同じ大きさになる様にデータベースの初期シグナチャーファイルをシステムに入れる事ができる。これには二つの技術がある。

(1) サブセットの選択は、ハッシュ機能を施す事により作成することができる。この無作為選択は、データベースに何れか新しい^{ワード}単語が加えられた時に何れかのサブセットにほぼ同一の選択機会を与えるのに寄与している。しかしながら(既に使用された)同じキーワードを必要とする新情報の追加は、いくつかのワードが繰り返し多用される場合は、特定のサブセットが他より早く成長する事がある。

(2) より優れた手段はキーワード辞書を用いる

もし700メガバイトのデータベースが70メガバイトの大型非細分化シグナチャーファイルを有しているとすれば、先の例に当てはめると細分化後のシグナチャーファイルは33%増しの993メガバイトとなる。この細分化されたファイルが256サブセットに分散されると各サブセットは約0.36メガバイトの容量となる。毎秒1メガバイトの転送速度では、平均的サブセットは0.36秒で転送されることができる。

シグナチャーサブセットは、走査中にディスクのヘッドの移動が必要でなければ最も効率良く走査させる事が出来る。こうする為には、新情報の追加に伴って増加してもサブセット情報ディスクの相隣接するセクタ内に納められていなければならない。これはカンセキューティブサブセット間に空スペースを設ける事で達成される。この空スペースはデータベースが段々増大するにつれて詰まって来る。データ記憶手段がディスクの場合、データ処理手段はディスクから一つ以上のサブセットを取り出すためにスキナと共同して使用され

事である。この方法は、データベースのテキストの二回走査を必要とする。一回目の走査で、システムは各々のキーワードを含む文書数を記憶しながらデータベース中の全ての別個のキーワードのリストを作成する。この文書の数は各キーワードにつき一つのサブセット内のワードシグナチャーグループの総記憶容量の推定に役立つ。キーワードは次に文書の数の小さい方から大きい方へと記憶される。すると、シグナチャーサブセットの長さがほぼ等しくなるように各キーワードのワードシグナチャーにサブセットを割当てればよい。この割当ては、キーワード辞書に書き取っておくのでデータベースの二回目の走査の時、システムはこのサブセットの割当てを物理ワードシグナチャーに連絡させる事により論理シグナチャーを組立てる。論理ワードシグナチャーと相応するキーワードの関連性をキーワード辞書に納める時、それらの論理ワードシグナチャー内のサブセット識別フィールドの値は、全てのサブセットの長さが同じになるように選ばれる。物理ワードシグナチャー

一は、論理ワードシグナチャーと相応するキーワード間に一対一の対応がとれる様に選択される。あるキーワードに対する物理ワードシグナチャーは、一度システムがどのキーワードがどの特定のサブセットに結びついているかを知ったならば、唯一無二的に定義する事ができる。これは、そのサブセットに割当てられた個別のキーワードに連続整数で番号をつける事により達成される。これらの番号が、物理ワードシグナチャーとなる。

アプリケーションの如何によって、システムは使用者に対して異なった応答をする。特定のクイリーキーワードがシステムに供給されると、システムはそのクイリーキーワードの物理ワードシグナチャーに一致する物理ワードシグナチャーを含んだサブセットのみを搜索する。サブセットは、そのクイリーキーワードを含んだデータベースの全ての文書の文書識別記号を含む。文書識別記号を用いる事により、全てのデータベースを走査することなくデータベースから文書を検索することができる。

ドレスとして用いられ、この場合、 n は8から20の整数値の範囲である。各ロケーションは1ビットを有し、スキャナはクイリーに対しRAMがそのワードシグナチャーは無関係であると指定した場合にそのワードシグナチャーを無視するように制御されている。そうでない場合は、その物理ワードシグナチャーをFIFOに入れる。RAMのビットロケーションに相補数ビットの値が納められている時は、RAMはシグナチャーファイルを格納しているディスクの転送速度と同じ処理速度で所要のワードシグナチャーを将来の参考として選別する。

次に述べる実施例では、物理ワードシグナチャーには15ビットの長さが用いられている。15ビットの物理ワードシグナチャーは、中央処理装置(以下CPUと記す)で動作しているソフトウェアによって前もって定義された内容を保持する32K×1に相当するRAMのアドレスに用いられる。もしアクセスされたビットロケーションに、この物理ワードシグナチャーが特定のクイリーによっ

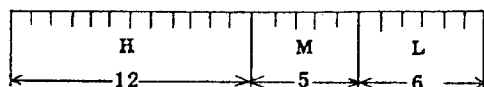
シグナチャーファイルのサブセットをソフトウェアで走査する一方、データ処理手段や中央処理装置等のリソースを、クイリーレセプション、クイリー解析、使用者のインタラクション及び入力動作の管理等のリアルタイム機能に充当するのが、より効果的である。ソフトウェアの代わりに、走査にはハードウェアモジュール(即ちスキャナ)を用いる事も可能であり、このハードウェアは予め決められた物理ワードシグナチャーを搜索する光ディスクからのデータストリームの発出を走査する。スキャナモジュールは、SCSI(スモールコンピュータシステムインターフェイス)プロトコール等のディスク転送プロトコールを受取るように設計されており、光ディスクとデータ処理装置のディスクインターフェイス間の転送を「聴取」する装置として本質的に作動する。

物理ワードシグナチャーに用いられるビット数は、システムの要求に依存するある幅の範囲で可能な値をとる。 n ビットの物理ワードシグナチャーは、RAMに於いて2の n 乗のロケーションのA

て追求されていない事を指示する2進数の値が含まれている時は、この物理ワードシグナチャーは無視される。もしアクセスされたビットロケーションに、上記相補数の値が含まれていると、その物理ワードシグナチャーと以下の文書識別記号は、CPUによって将来の参考用としてファーストインファーストアウト(FIFO)バッファに入力される。高速RAMを使用する事により、シグナチャーファイルを保持しているディスクの転送速度に見合った処理速度で所要のワードシグナチャーを選別する事が可能になる。特定の物理ワードシグナチャーが無視されると、システムは次の物理ワードシグナチャーを調べる。論理ワードシグナチャーで用いられる好ましいビット数を m ビット数を m ビット長とすると、 m は8から32までの範囲にある。

シグナチャーファイルによって利用されるディスクスペースを最小にする為、文書識別記号の表示値に平均して必要なスペースを最少にする必要がある。これを達成する一つの方法は、シグナチャー

ーファイル作成時に文書番号を三つのフィールドから成る一つの値として取り扱う事であり、その三つのフィールドは上位フィールドのラベルをH、中位フィールドのラベルをM、そして低位フィールドのラベルをLとして次の様に表される。



2 3 の文書番号

しかしながら、文書番号は必ずしも連続通し番号とは限らないが、特定のサブセット以内では番号の増加する方向で現れる。従って、H又はMフィールドは、低位フィールドが最後の表示値以後にゼロの値を通過した時のみ次の文書識別記号の表示値に加えられる。走査動作時に、Hフィールドは常にその発生時にFIFOに入力される。Mフィールドが上記ストリームに現れた時、これはMREGレジスタと称されるレジスタに将来使用するため与えられる。各文書のワードシグナチャーのグル

れはワードシグナチャーストリームのソースの確認用として各セクタの始めに埋め込まれている。

ープは、Lフィールドを含む単独のバイトによって終わらせることができる。Mフィールドは、その前の時点でのLフィールドの値に対応していた値より増加した時のみ、Lフィールドの前の位置に挿入される。文書識別記号の表示値はFIFOに入力されるべき時、最後に現れたMフィールドの値（先にMREGに与えられた値）と組み合わせられる。FIFOの出力を処理しているソフトウェアは、ストリームに最後に現れたHフィールドと一組のM、Lフィールドを組み合わせる。このようにして、文書への関連づけは平均して一文書当たり1バイトよりやや大きい程度の小さなスペース負担で処理され得る。文書識別記号の表示値が「Lフィールド」のみあるいは「M及びLフィールド」のみから成るとしても実際の文書識別記号は三つのフィールド全てを有する。

表1は、シグナチャーファイルサブセットにおけるバイト及びワードのエンコーディングの例を表わす。表1では、LSSA及びHSSAは、夫々シグナチャーセクタアドレスの低位及び高位であり、そ

表 1

サブセットのタイプ	形 式	ビット長	動 作
ワードシグナチャー	 WS=15ビットワードシグナチャー	16	マッチがとれた場合FIFOに入力
文書番号 Lフィールド	 L=文書番号のビット0からビット5	8	このグループのシグナチャーが検出された時MREGと組合せてFIFOに入力する
文書番号 Mフィールド	 M=文書番号のビット6から10	8	MREGに貯える
文書番号 Hフィールド	 H=文書番号のビット11から22	16	FIFOに入力
シグナチャーセクタ アドレスの低位 部分	 LSSA=セクタアドレスのビット0から10	16	FIFOに入力
シグナチャーセクタ アドレスの高位 部分	 HSSA=セクタアドレスのビット11から21	16	FIFOに入力

もっと一般的に言えば、シグナチャーファイルの作成時に各サブセットを作る為に一連のグループが生成される。各グループは一連の物理ワードシグナチャーを持ち、文書識別記号の表示値で終わらせられる。実際の文書識別記号は高、中、低位のフィールドを持つ。あるサブセット内の最初のグループの文書識別記号の表示値は、事実、高、中、低位のフィールドを持つ。次のグループから、文書識別記号の表示値は必ず低位のフィールドを保有するが、その直前のグループで用いられた文書識別記号との相異に対応すべく、必要に応じて中位又は高位のフィールドを持つことになる。文書識別記号は数の増加する方向に設定されている。従って、特定のグループ用のある文書識別記号の表示値が低位のフィールドのみを有するとしても、実際のそのグループ用の文書識別記号は高、中、低位のフィールドを有する事になる。これは一つの変形として、もしスペースを節約しなくても良い場合は全ての文書識別記号の表示値は実際三つのフィールドを持つ事が出来る。

ませ得ることができる。

4. 図面の簡単な説明

第1図は、本発明のデータ記憶及び検索方法及びスキヤナの概略を示すブロック構成図、第2図はデータ記憶及び検索用のデータ処理手段を示すブロック図である。

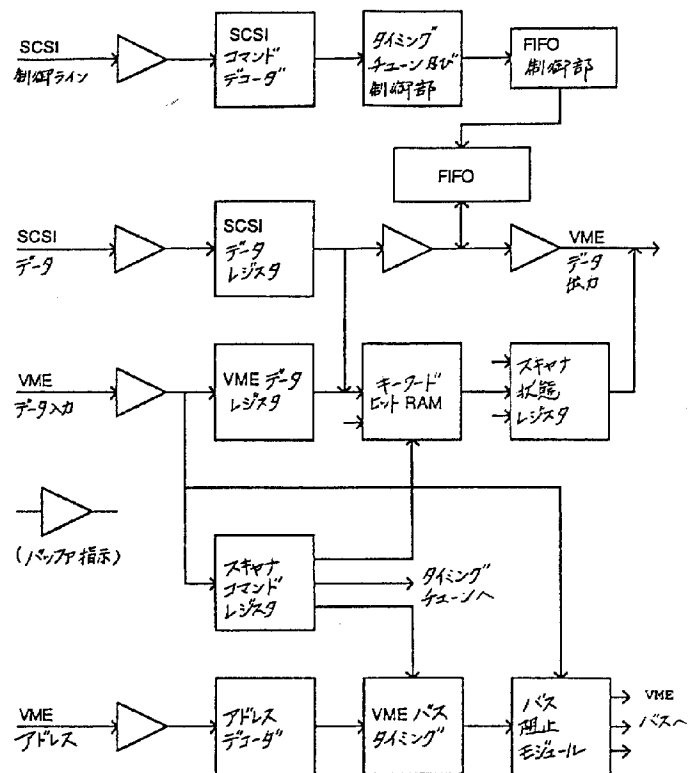
出願人代理人 弁理士 鈴 江 武 彦

データ記憶手段からのデータストリームをスキヤナを用いて走査する事によりデータ記憶領域から情報を検索する場合、スキヤナは高位フィールドを有する全ての文書識別記号をFIFOに納める。スキヤナは将来必要になる時の為に、データストリームで最後に遭遇した中位のフィールドをレジスタに貯えておく。クイリーキーワードの物理ワードシグナチャーと特定のグループに於ける物理ワードシグナチャーの間に一致がとれた事をRAMが指示した時のみ、最後に遭遇した中位のフィールドがその特定のグループを終わらせる低位のフィールドと共にFIFOに挿入される。

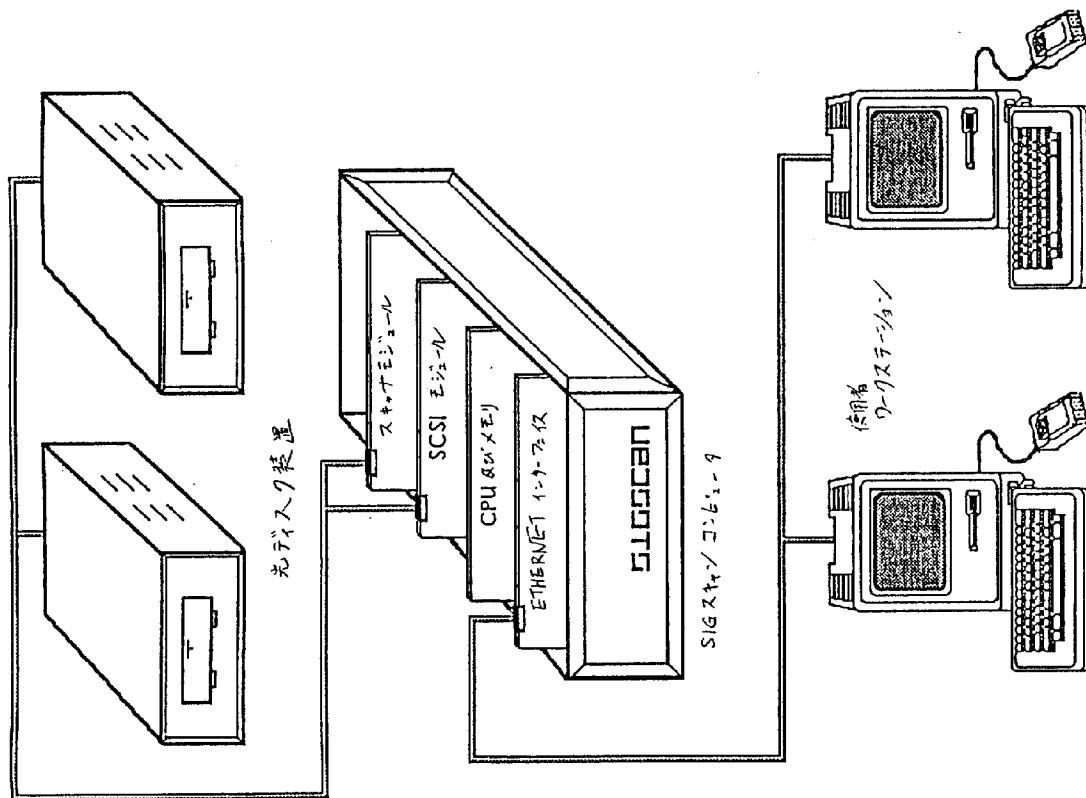
尚、本発明は、上述した実施例に限られるものではなく、本発明の要旨を変えない範囲において、多くの変型が容易に考察されることは勿論である。
〔発明の効果〕

以上のように、本発明によれば、シグナチャーファイルの検索が一回のみで十分であり、且つ特定のクイリーキーワードに対する応答としての検索が、単にシグナチャーファイルの一部のみで済

図面の浄書(内容に変更なし)



第2図



第 1 図

手 続 補 正 書 (方式)

平成 1 年 5 月 30 日
1. 5. 30

特許庁長官 吉 田 文 毅 殿

1. 事件の表示

特願平 1 - 0 1 1 7 5 2 号

2. 発明の名称

データ記憶及び検索方法及びスキャナ

3. 補正をする者

事件との関係 特許出願人

氏名 フォーブス・ジェイ・パーコブスキ (ほか 1 名)

4. 代理人

住所 東京都千代田区霞が関 3 丁目 7 番 2 号

〒100 電話 03(502)3181 (大代表)

氏名 (5847) 弁理士 鈴 江 武 彦



5. 補正命令の日付

平成 1 年 4 月 2 5 日

6. 補正の対象

図面 (第 2 図)

7. 補正の内容 別紙の通り

図面の添付 (内容に変更なし)

